DOCUMENT RESUME

ED 350 355 TM 019 138

AUTHOR Yamamoto, Kentaro; Kulick, Edward
TITLE An Information-Based Approach To Monitoring Content Validity and Determining the Relative Value of Polytomous and Dichotomous Items.
PUB DATE [92]
NOTE 35p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Content Validity; Elementary Secondary Education; *Field Tests; Mathematical Models; *Sciences; Test Construction; Test Content; *Test Items
IDENTIFIERS Dichotomous Variables; *National Assessment of Educational Progress; Polytomous Items

ABSTRACT

Test items are designed to be representative of the subject areas that they measure and to reflect the importance of specific domains or item types within those subject areas. Content validity is achieved by content specification and number of items in each content domain included in the design of the test. However, largely due to the normal attrition of less desirable items during the field test phase, the averaged proportional domain information of the 1990 National Assessment of Educational Progress (NAEP) reading assessment deviated from the original design. The relative value of polytomous and dichotomous items in a 1990 NAEP science assessment booklet was evaluated. The current study included 1,248 students from the NAEP grade 8/age 13 years sample who were administered booklet 20, which contains 3 science blocks. Results show that the polytomous items were only as informative as dichotomous items in regard to both the model-based relative information values as well as the reduction in the variance of posterior proficiency distribution. It is contended that in the future, there should be a greater concern that the reporting score reflects the content validity of the instrument specifications. The components of information should be monitored routinely by the item response models or by the content, and the reduction of the posterior variance should be monitored through additional items. Six tables and four graphs illustrate the analysis. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED350355

An Information-Based Approach to Monitoring Content Validity and Determining the Relative Value of Polytomous and Dichotomous Items

by
Kentaro Yamamoto and Edward Kulick
Educational Testing Service

Abstract

Test items are designed to be representative of the subject areas that they measure and reflect the importance of specific domains or item types within those subject areas. The content validity is achieved by the content specification and number of items in each content domain included in the design of the test. However, by largely due to the normal attrition of less desirable items during the field test phase, the averaged proportional domain information of the 1990 NAEP Reading assessment deviated from the original design. Relative value of polytomous and dichotomous items in a 1990 NAEP Science assessment booklet was evaluated. It was found that polytomous items were only as informative as dichotomous items in regard to both the model based relative information values as well as the reduction in the variance of posterior proficiency distribution.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
KENTARO YAMAMOTO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM019138

BEST COPY AVAILABLE

An Information-Based Approach to Monitoring Content Validity and Determining the Relative Value of Polytomous and Dichotomous Items

Introduction

Test items are designed to be representative of the subject areas that they measure and reflect the importance of specific domains within those subject areas. Such representations are achieved by the content specification and number of items in each content domain. By so doing, the content validity is established. Often the relative importance and appropriateness of the domains within a subject area differ among themselves depending upon characteristics of the intended examinees.

A calculus test would be appropriate for some college students, however, the same test would be inappropriate for elementary schoold children or some other college students without a previous calculus course works. Any results from such a test on the inappropriate samples would be invalid in regard to the ability of the examinees. The domains of a subject area need to be represented differentially in accordance with the characteristics of examinees in order to assure the validity of the measurement.

Application of this notion of content validity should not be limited to the dichotomous items but should extend to polytomous items as well. Often the polytomous item response model may be applied to constructed response items (open-ended items), in order to measure unique qualities or to assess performances in a more "authentic" situation. It is common when assembling a measurement instrument to select items differing in response mode, e.g., multiple

choice and constructed response. In addition, one can assign differential importance to different types of response modes.

Monitoring Content Validity

Let us first consider a hypothetical situation as an example, before we discuss the reading scale based on the 1990 NAEP reading assessment. Suppose two domains define a subject area, and both are equally important. Content validity of the assessment instrument was established by including an equal number of items from each domain. After item parameter calibration, all items from the first domain "A" were found to be very informative with regards to the proficiency scale of theta (very steep slope parameters) compared to items from the other domain "B" (very flat slope parameters). As a result, proficiency estimates can be completely dominated by the results from domain A, regardless of the fact that both domains contain an equal number of items, in accordance with the original definition of the subject area. In fact, such a proficiency scale would represent a domain composition vastly different from the originally defined subject area. This example illustrates why confirmation of content validity through comparisons of design domain specifications and the calibrated domain representation should be made. At the present time, this confirmation is rarely, if ever, performed. Typical attempts to verify an instrument's content validity end with a comparison of the number of items in each domain. Our hypothetical example shows why this is not a sufficient effort.

The development of the 1990 National Assessment of Educational Progress (NAEP) assessment of reading was documented in the NAEP publication *Reading Objectives, 1990 Assessment*. That document states that in regard to the

placement of cognitive reading items into text categories,

> The assessment Development Panel considered these materials as belonging to three categories: informational text, literary text, and documents.... NAEP staff and consultants perceived that most passages fit satisfactorily into one category or another. The Assessment Development Panel chose to highlight informational text, literary text, and documents because these categories represent the types of materials that students commonly encounter in and out of school and are expected to be capable of reading.

That statement provides the basis for the content validity of the reading assessment.

The 1990 NAEP sample consisted of nationally representative samples from three grade levels, specifically, 4, 8, and 12. Subject area committees comprised of experts in reading research and education were assembled to determine the relative importance of these three domains for each of the three grade levels. The following table shows the relative importance assigned by these committees to the three reading domains for each grade.

---

insert Table 1 about here

Preliminary test forms were assembled using items that best characterized each of the three domains and in proportions that closely matched the relative importance of the domains in each grade level. These items were administered to a field test sample to ensure their appropriateness for the assessed grade level and to gauge the goodness of fit of the IRT model to the data. This process, however, resulted in the attrition of more items between the field test and the test's final form than was previously expected. In addition, relatively more items from the smallest domain (Documents) were identified as not appropriate. As a result, the composition of the final assessment instrument was noticeably

different from the committees' original domain specifications. The following table shows the proportion of items actually included in the final analysis.

---

insert Table 2 about here

After these items were administered and response data appropriately scored, the unidimensional 3pl IRT model was used to calibrate parameters of items from three domains simultaneously. It is assumed that the resultant estimates of proficiency would reflect the relative importance assigned to the three domains regardless of the dissimilarity to the proportions of items actually administered. The item calibration method being applied here maximizes the reliability of scores without paying attention to content domain.

As noted earlier, the 1990 NAEP reading scale had three domains, and their relative importance and the numbers of items in each were presented in Tables 1 and 2. The test information by domain, i.e., the sum of item information in a domain, was calculated by the following formula in order to estimate how much information from each domain is reflected in the estimated proficiency value:

$$I(\theta, M=m) = \sum_{I_m=1}^{N_m} \frac{P'^{2}_{i_m}}{P_{i_m} Q_{i_m}}$$

where m identifies the domain, P_i is the conventional 3-parameter logistic (3PL) IRT model, Q_i is $(1.0-P_i)$, and P'_i is the first derivative with respect to θ.

The domain test information can be standardized by the total test

information at each theta value in order to evaluate the conditional proportion of domain test information given theta, as follows:

$$I_s(\theta, m) = \frac{I(\theta, m)}{\sum_{m=1}^{M} I(\theta, m)}$$

Figures 1, 2, and 3 present the proportional domain test information for the 1990 NAEP reading assessment of grades 4, 8, and 12, respectively. For grade 4, the Informational Text domain contributed very little to the proficiency estimation, compared to the contribution of the Literary Text domain in the lower proficiency range. However, that relationship reverses at the higher proficiency range. The Documents domain consistently contributes much less to the total information for the entire range of theta than its designed importance would suggest. Some interactions of proportional domain test information with proficiency value were observed at all grade levels. For grade 12, the direction of the interaction between the Informational Text domain information and proficiency value was opposite to that observed in the other two grade levels.

Averaged information can be obtained by integrating $I(\theta,m)\cdot f(\theta)$ and can be expressed as follows:

$$E(I) = \int_{\theta} I(\theta, m) \cdot f(\theta)\, d\theta .$$

In the same way that the conditional domain information was standardized, the averaged information values can also be proportionalized. The averaged proportional domain information can be used to assess the relative contribution of each domain, while considering the relevant proficiency distribution. That

is to say, conditional domain information in the proficiency range most attained by examinees is weighted more heavily than information from less frequently attained ranges. The following table presents proportional averaged domain test information. Although these values reasonably approximate the proportion of items in each domain, they grossly deviate from the original design of relative domain importance.

---

insert Table 3 about here

The Informational Text domain was represented in the proficiency scale fairly closely to the original design for grade 4. However, for grades 8 and 12, the domain was overrepresented. The Literary Text domain was overrepresented for grade 4, underrepresented for grade 8, and fairly represented for grade 12. The Document domain was consistently underrepresented by one third for all grade levels.

Due to the real problems any survey research project would face, the proportion of items from each domain can often be difficult to control closely as evidenced in the NAEP reading data. The result can be differences between the mix of items in the actual assessment and the framework specifications. Consequently, alternative methods to those that completely depend on the number of items warrant examination.

Several alternatives are currently available. For example, in the 1990 NAEP mathematics assessment, there were five domains of mathematics. In a similar fashion to the reading scale, the mathematics subject area committee established the relative importance of domains for each grade level and

incorporated them in the mathematics objectives. The mathematics items were developed according to guidelines reflecting these objectives. Subsequently, they were field tested just like the reading scale items. The difference exists in the treatment of domains. Each domain was treated as a separate subscale in the mathematics scaling, while all domains were treated as one scale in the reading analysis. By maintaining separate subscales, the intended content domain representation can be assured by assigning proper weights to each subscale score and producing a composite score, as was done in the NAEP mathematics scale.

Multidimensionality of the domains in a subject area can cause a relatively greater attrition of items from the smaller scales due to a lack of fit of the IRT model. It is evidenced on many occasions that scales containing the largest number of items tend to dominate the calibration process of the IRT model. Consequently, items from a small scale are more frequently identified as poorly fitting the measurement model. This situation is exaggarated when a scale actually represents a unique dimension. In the 1992 NAEP analysis, the reading scale is scheduled to have subscales following the convention of mathematics and science analyses. This method should result in proficiency estimates that reflect the content validity described in the reading objectives more readily and accurately.

In addition to the aforementioned procedure to reflect the content domain weights, at least two methods can be discussed. The first is to use a multidimensional IRT model when parameters of items from several domains are calibrated together. This method would enable us to use the covariance of dimensions due to domain differences. This procedure would require a far greater amount of information in terms of the number of examinees, the number of items, and computer resources. What this method cannot do is to incorporate content

domain weights at the time of item calibration. However, after the item calibration is completed, a composite score can be calculated based upon the domain scores in a manner that reflects content domain weight specifications. The second, less desirable procedure, is to apply domain weights after a unidimensional IRT model has been applied. In other words, use such weights only to produce scores. Such a procedure may be described as

$$\log L_v(x,y|\theta) = w_1 \cdot \log l(x|\theta) + w_2 \cdot \log l(y|\theta)$$

where w_1 is the weight for the domain x and w_2 is the weight for the domain y. Undoubtedly, this formulation would produce proficiency estimates more in line with the content validity which was designed, but in return we would have less reliable estimates of ability (because they would be based on fewer items) on the reporting scale. It is uncertain whether one method is uniformly better than the other. However, it is clear that in the future there should be a greater concern that the reporting score reflect the content validity of the instrument specifications.

Determining the Relative Value of Polytomous and Dichotomous Items

The aforementioned monitoring procedure can also be used to determine the relative value of polytomous and dichotomous items when both item types are included on the same test. A widely accepted notion of the benefit from using polytomous items over dichotomous items is that the polytomous items are more informative for a wider range of ability. This stems from the belief that we learn about one's ability not only on the basis of a correct or incorrect

response, but also from one's choice of incorrect response. Before gaining general acceptance, such claims need to be validated using real data.

If polytomous items truly enjoy a 'more information per item' advantage over dichotomous items, then this information needs to be used when creating test specifications regarding item types. Merely assuming that equal numbers of items of each type will yield equal amounts of information from both item types would be erroneous. If polytomous items actually provide twice as much information as dichotomous items, then on a test comprised of equal numbers of both item types 67% of the information, as reflected in the final ability estimates, may be expected to come from the set of polytomous items.

In addition to quantitative gains in information obtained from a polytomous item response model, it is also claimed that "constructed response" or "authentic" items measure somewhat qualitatively different aspects of ability, which cannot be measured by traditional dichotomous items. If polytomous items truly tap an underlying construct different from that measured by dichotomous items then the simultaneous estimation of both types of items is in violation of the model's assumption of an underlying unidimensional construct. On the other hand, if both types of items are on a unidimensional ability axis so that simultaneous estimation is appropriate, the claim of unique measurement properties for polytomous items cannot be true.

Until recently, it was not possible to estimate the parameters of polytomous items and dichotomous items simultaneously due to a lack of appropriate computer programs. This is no longer a restriction, and one currently available computer program, PARSCALE (Muraki & Bock, 1991), was used for the following analyses. This program enables the user to estimate parameters of various combinations of item types simultaneously. The measurement models

used in PARSCALE are the three-parameter logistic IRT model (Lord & Novick, 1968) and the generalized partial credit model (Muraki, 1992) based on the Master's partial credit model. The response functions of these two models are as follows:

$$P_j(\theta) = c_j + \frac{1-c_j}{1+\exp[-Da_j(\theta-b_j)]}$$

where a_j, b_j, and c_j are the 3PL IRT item parameters of item j.

$$P_{jk|k-1,k}(\theta) = \frac{\exp[Da_j(\theta-b_{jk})]}{1+\exp[Da_j(\theta-b_{jk})]}$$

where k=2, 3,...m_j, and both a_j and b_{jk} are the generalized partial credit model parameters.

The 1990 NAEP science assessment was administered to a nationally representative sample of three student cohorts: students who were either in the fourth grade or 9 years old, students who were either in the eighth grade or 13 years old, and students who were either in the twelfth grade or 17 years old. The pool of items used in the 1990 science assessment contained a range of open-ended and multiple-choice questions that were developed in agreement with a set of objectives documented in *Science Objectives: 1990 Assessment* (NAEP, 1989b). The science items were put together into seven distinct blocks for each grade/age cohort separately. Then these seven blocks were combined together to form seven books, each block appearing in three different within-book positions (first, second or third in a booklet) exactly once.

The current study included all of the 1,248 students from the NAEP grade 8/age 13 sample who were administered booklet 20. Booklet 20 contained three science blocks, SH, SI, and SD, respectively. Block SH consists of mainly open-ended items, many of which required constructed responses from the examinee.

(Although two other booklets included block SH, they were not selected for analysis to ensure that every student was administered the identical item set.) The booklet included 44 dichotomously scored items, mostly in multiple-choice format, and 12 polytomously scored items. Note that some open-ended items were scored simply right or wrong, and hence included in the dichotomous item set. Since blocks are separately and equally timed, the influence of a poor performance in a previous block was expected to be minimal for the following block.

Table 4 displays some descriptive statistics for each of the three blocks. The open-ended items were dichotomized in order to be included in the calculation of block statistics. Note that the number of items in block SD (all dichotomous items) is 26 - this is nearly double the number of open-ended items (14, five of which were scored dichotomously) in block SH. So for a fixed amount of testing time, many more multiple choice items than polytomous items were administered. Differences in economy and efficiency in the administration of certain item types can be justified if they provide correspondingly more information.

---

insert Table 4 about here

For the subsequent analyses, the model parameters were estimated using three different combinations of items, 1) dichotomous items only, 2) polytomous items only, and 3) both dichotomous and polytomous items together. The NAEP analyses employed plausible value methodology in order to estimate the proficiency distribution of subgroups. However, the plausible value is not the

best estimate for an individual examinee. Also the methodology requires fairly large numbers of examinees for a proper application. We used the Marginal Maximum Likelihood method based on the posterior distribution of an examinee to estimate mean and variance of proficiency, and the posterior distributions of every examinee were combined to estimate the population distribution.

We examined the amount of information reflected in the final ability estimates contributed from each of the two item types using two different methods - a classical test theory approach and a model-based method. The first method is based on comparisons of the total posterior ability distributions and the posterior error variance. When the number of items on a test is increased, it is expected to increase the amount of information regarding the student's ability. As test information increases, there should be a corresponding decrease in the variance of an examinee's posterior ability distribution. The total population proficiency variance can be divided into two parts, one is between point estimates of examinees' proficiency and the other is averaged uncertainty within an examinee. This division is analogous to the variance components used in the analysis of variance. The variance between examinees can be estimated by calculating the variance of expected a posteriori estimates (EAP) of proficiencies. The EAP estimate of each examinee's proficiency can be expressed as,

$$\theta_j = E(\theta|\underline{X}_j) = \int \theta \cdot \frac{P(\underline{X}_j|\theta)}{P(\underline{X}_j)} d\theta .$$

The uncertainty (within examinee variance) of the expected a posterior can be calculated as

$$Var(\theta_j) = \int (\theta - \theta_j)^2 \frac{P(\underline{X}_j|\theta)}{P(\underline{X}_j)} d\theta .$$

Hence, the population variance can be written as the sum of two components of variance, namely the variance of EAP estimates and the averaged variance of examinee proficiency distributions,

$$Var_{pop} = Var(\theta) + \frac{\sum_j Var(\theta_j)}{N}$$

The proportional variance of averaged uncertainty of EAP in reference to the total variance using only dichotomous items was 11.98%. However, when polytomous items were included, the ratio was decreased to 9.72 %. This is nearly a 19% reduction in the posterior variance correspondence, not quite matching the increase of items from 44 to 56 (27.3%). Some rough estimates of test information can be made using the following relationship.

$$information = \frac{Var_{population}}{Var_{error}}$$

The averaged information per item for all three analyses are reported in Table 5.

---

insert Table 5 about here

The model-based test information formula presented in Lord and Novick (1968) for the 3PL IRT model is

$$I(\theta) = \sum_{j}^{n} \frac{[\frac{\partial}{\partial\theta} P_j(\theta)]^2}{P_j(\theta) \cdot Q_j(\theta)} \qquad \text{(3PL IRT model)}$$

where $P_j(\theta)$ is the conditional probability of making a correct response on item j given proficiency θ. Samejima (1974) presented the following comparable formula for the polytomous item response model, which sums the contribution of each category, k, weighted by the probability of that category,

$$I(\theta) = \sum_{j}^{n} \sum_{k}^{m} P_{jk}(\theta) \, [-\frac{\partial}{\partial\theta^2} P_{jk}(\theta)] \qquad \text{Polytomous model}$$

where $P_{jk}(\theta)$ is the conditional probability of being in response category k on item j given proficiency θ. Following the same notion used in the previous section on dichotomous items, the proportional information at a given ability level was calculated and plotted (Figure 4). This figure clearly indicates that our polytomous items were more informative relative to the total number of items in both extreme ability ranges, roughly corresponding to theta values below -1.0 and above 2.0. However, in the more populous range, theta values between -1.0

and 2.0, polytomous items were slightly less informative than dichotomous items, relative to the total number of items. Consequently, ability estimates based on the simultaneous use of the two models have differential representation in terms of item types depending upon ability level. In addition, our polytomous items were not as informative as our dichotomous items for the most populous range.

In order to summarize the results of Figure 4, averaged information was calculated following the formulation presented earlier. Here in place of $f(\theta)$ an estimated proficiency distribution was used and a numerical integration method was applied. If in fact both polytomous items and dichotomous items were measuring the same ability, the test information from the entire 56 items would be equal to the sum of information obtained separately.

---

insert Table 6 about here

The table indicates that the amount of information when both types of items (polytomous and dichotomous) were used is 93% of the sum of the two separate information components, a decrease equivalent to about 4.5 dichotomous items. In other words, if we assume that the average information of .232 for the dichotomous items holds, we could have had the same accuracy of ability estimates by adding 6 more dichotomous items instead of 12 more polytomous items.

Discussion

This paper presented a construct to monitor content validity in a way that really matters, not in the number of items but in the ability estimates that each content domain should uniquely represent. However, it should be noted that not every interaction of content representation and ability is wrong or should be

avoided. Sometimes it is desirable that one domain should contribute more to the ability estimate than others in a higher ability range. If mathematics were unidimensional, an Algebra and Function domain would be more informative for the higher ability examinees than items in a Numbers and Operations domain. We believe that test makers should be cognizant of the fact that content validity cannot be assured by the number of items alone. We focused only on the content validity in this paper, however, the notions of test validity remain extremely important and readers should referr to the extensive discussion on the subject by Messick (1989).

In regard to the simultaneous estimation of dichotomous and polytomous item response models, this current study is too limited to reach conclusions on the relative value of polytomous items. It is clear that among polytomous items some are more informative than others, just as among dichotomous items some items are more informative than others. Moreover, some polytomous items are not as informative as some dichotomous items. Describing a test in terms of the proportion of items either by content domain or by item types is grossly inadequate. What is not clear is the specific characteristics of informative polytomous items, we can speak only in generality, such as informative polytomous items should elicit responses that correspond to the ability in order. In addition, scoring categories should be able to capture such pre-existing order. This type of within item correspondence to ability through ordered responses for the polytomous items provide additional constraints on the ability beyond dichotomous items. However, it may be said that if the goal of testing is to place every examinee in a continuum, then some polytomous items may not be suitable and dichotomous items do it better, cheaper, and faster. Under such conditions the inclusion of polytomous items may contribute only to increase the

face validity. However, some qualitative information may be measured only through polytomous items. In some cases dependence of various errors made on several items may be more informative in regard to examinees' solution strategies and/or a particular misunderstandings, and also such results are more useful for the learning than relative standing in the class room.

We recommend to monitor routinely the components of information by the item response models or by the content, and also to monitor reduction of the posterior variance through additional items.

References

Allen, N. (1992). *Data Analysis for the Science Assessment.*In E. G. Johnson & N. L. Allen (Eds.), *The NAEP 1990 techinical report* (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Donohugh, J. (1992). *Data Analysis for the Reading Assessment.* In E. G. Johnson & N. L. Allen (Eds.), *The NAEP 1990 techinical report* (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Lord, F. & Novick, M (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Masters, G. (1982). A Rasch model for partial credit scorign. *Psychometrika, 47,* 149-174

Messick, S. (1989). *Validity.* In R Linn (Ed.), *Educational Measurement.* New York, NY: American Council on Education, and Macmillan Publishing Co.

Muraki, E. (in press). A geraralized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement.*

Muraki, E., & Bock, D. (1991). *PARSCALE: Parameter scaling of rating data.* Chicago, ILL: Scientific Software, Inc.

National Assessment of Educational Progress. (1989). *Reading objectives, 1990 assessment.* (No. 21-R-10) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

National Assessment of Educational Progress. (1989). *Science objectives, 1990 assessment.* (No. 21-S-10) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111-121.

Table 1
Relative Importance (Proportions) of Three Reading Domains
by Grade

Domain	Grade 4	Grade 8	Grade 12
Informational Text	.40	.50	.55
Literary Text	.40	.30	.20
Documents	.20	.20	.25

Table 2
Proportion of Items in Three Reading Domains
by Grade

Domain	Grade 4	Grade 8	Grade 12
Informational Text	.42	.73	.68
Literary Text	.52	.16	.21
Documents	.06	.10	.12

Table 3
Proportional Averaged Domain Test Information by Grade

Domain	Grade 4	Grade 8	Grade 12
Informational Text	.37	.71	.69
Literary Text	.57	.21	.24
Documents	.06	.08	.07

Table 4
Descriptive Statistics for Item Blocks for Grade 8

Statistic	Block		
	SH	SI	SD
Total Number of Scaled Items	14	16	26
Number of Polytomously Scaled Open-Ended Items	9	3	0
Unweighted Sample Size	1233	1246	1245
Average Weighted Proportion Correct	.38	.58	.52
Averaged Weighted R-Biserial	.58	.57	.53
Weighted Alpha Reliability	.69	.69	.81
Weighted Proportion of Students Attempting Last Item	.87	.97	.92

Table 5

	Three estimation procedures		
	Dichotomous	Polytomous	Dichotomous + Polytomous
N. of items	44	12	56
Total Test Information	8.35	3.02	10.29
per item T.I.	.190	.254	.184

Table 6

	Three estimation procedures		
	Dichotomous Only	Polytomous Only	Dichotomous+ Polytomous
N. of items	44	12	56
model-based Test Information	10.23	2.44	11.70
per item T.I.	.232	.203	.209

Figure 1
Proportional Domain Test Information
NAEP 1990 Reading Data - Grade 4

Figure 2
Proportional Doamin Test Information
NAEP 1990 Reading Data - Grade 8

Figure 3
Proportional Domain Test Information
NAEP 1990 Reading Data - Grade 12

Figure 4
Proportional Test Information by Item Format
NAEP 1990 Science Data - Grade 8

FIGURE 1

PROPORTIONAL DOMAIN TEST INFORMATION

1990 NAEP READING DATA - GRADE 4

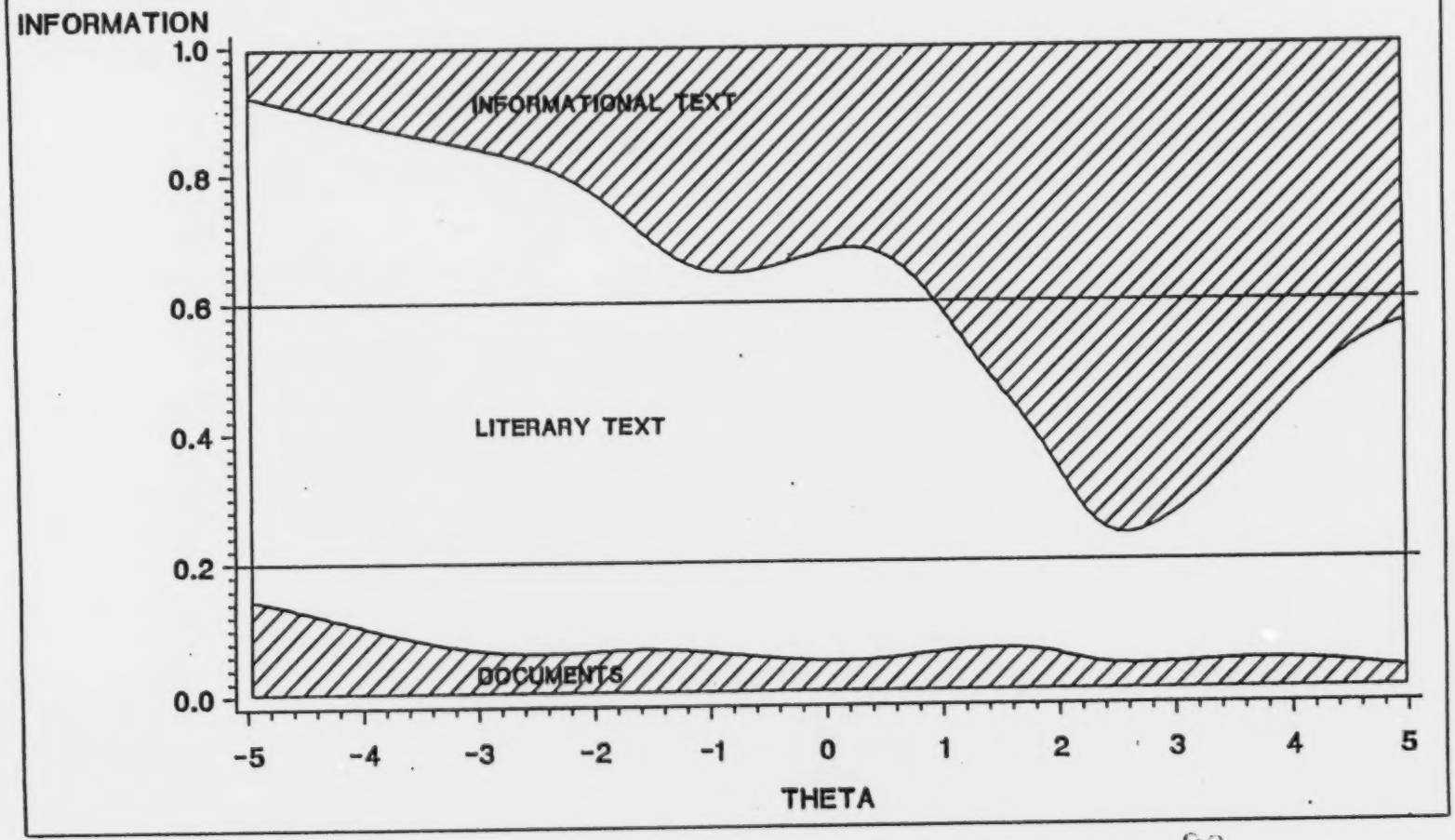

FIGURE 2

PROPORTIONAL DOMAIN TEST INFORMATION

NAEP 1990 READING DATA - GRADE 8

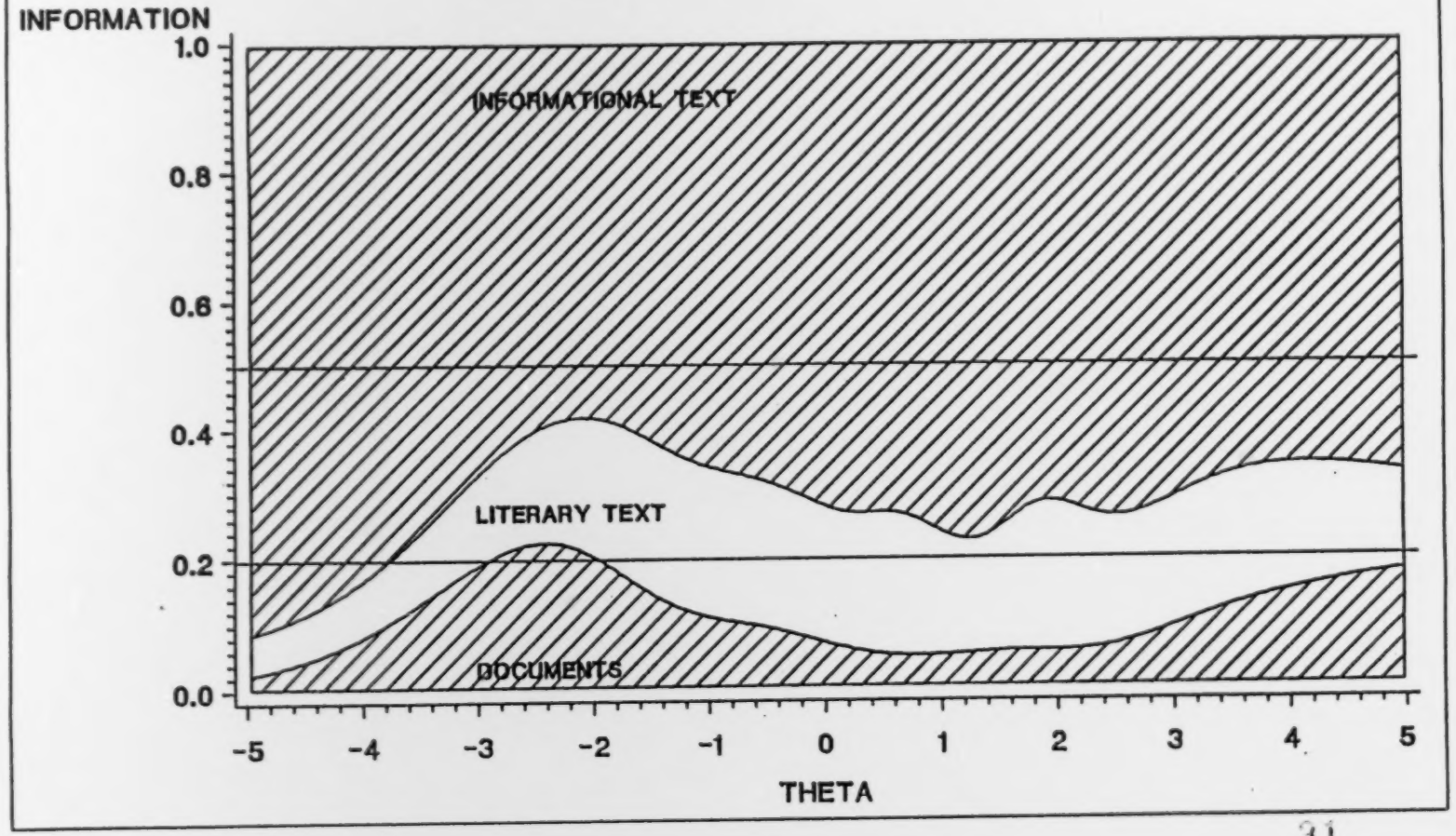

FIGURE 3
PROPORTIONAL DOMAIN TEST INFORMATION
NAEP 1990 READING DATA - GRADE 12

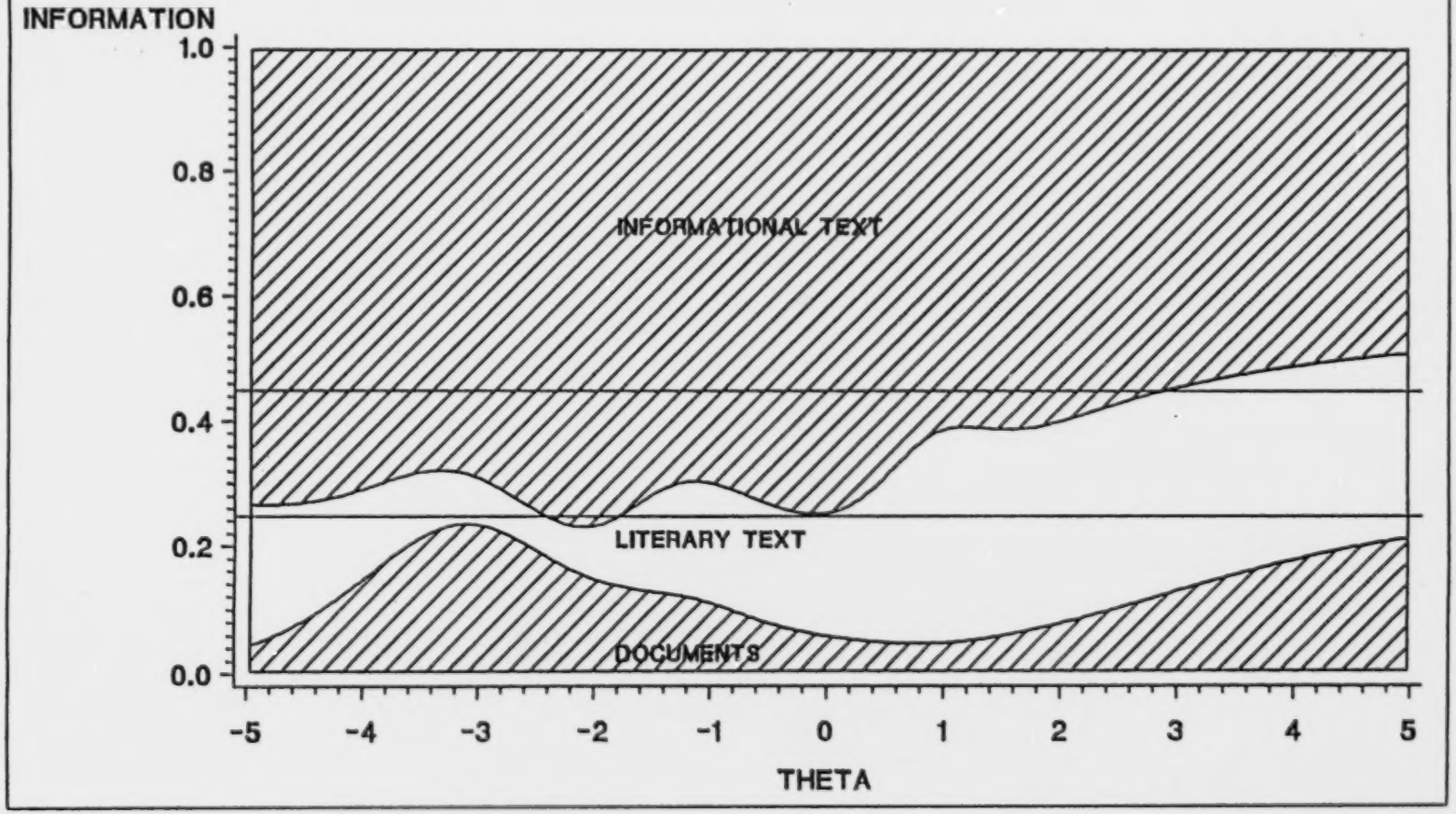

FIGURE 4
PROPORTIONAL TEST INFORMATION BY ITEM FORMAT
NAEP 1990 SCIENCE DATA - GRADE 8